Mixed Precision Customized for Discontinuous Galerkin Methods

Shivam Sundriyal, Markus Büttner, Vadym Aizinger
Chair of Scientific Computing, University of Bayreuth, Bayreuth, Germany

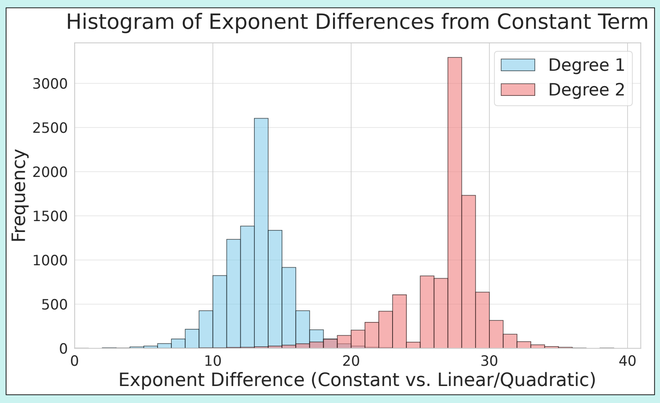
Hierarchical Basis Functions and Coefficient Decay

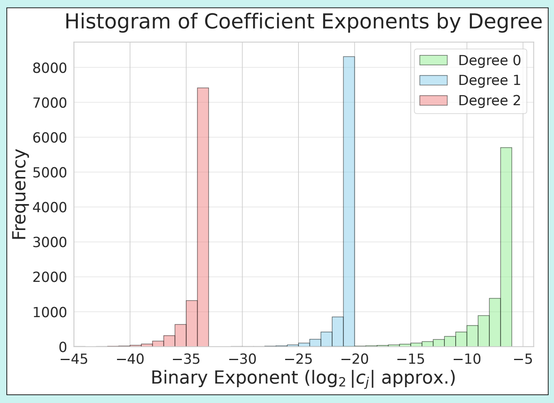
• Discontinuous Galerkin (DG) methods commonly use hierarchical basis functions, such as Legendre polynomials.

• The solution is expressed as: $u(x) \approx \sum_{j=0}^{p} c_j P_j\left(\frac{2}{h}x\right)$

• Coefficient trend[1]: $c_j \sim h^j$

• Implication:
Low-degree coefficients are large
High-degree coefficients are small → can be stored in reduced precision



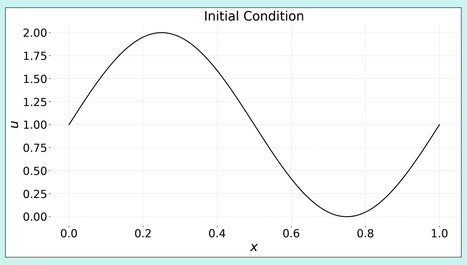


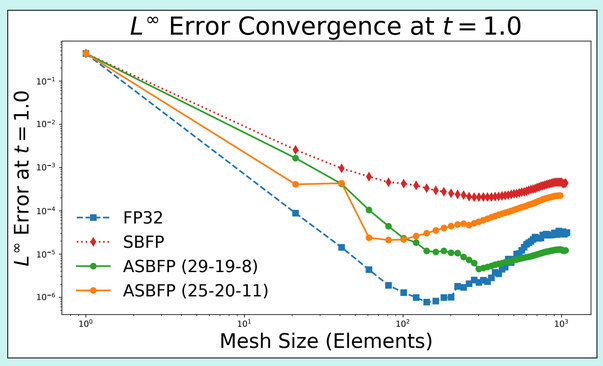
Problem Setup: Linear Transport Equation with DG

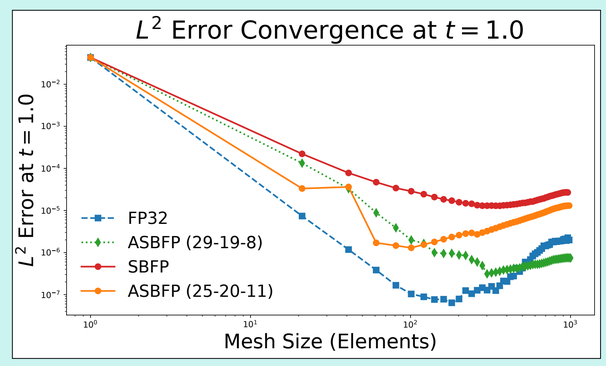
We demonstrate our Mixed Precision stratergy on the 1D linear transport equation:

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial u(x,t)}{\partial x} = 0, \quad x \in [0,1],$$

with periodic boundary conditions and the smooth initial condition : $u(x,0) = \sin(2\pi x) + 1.$

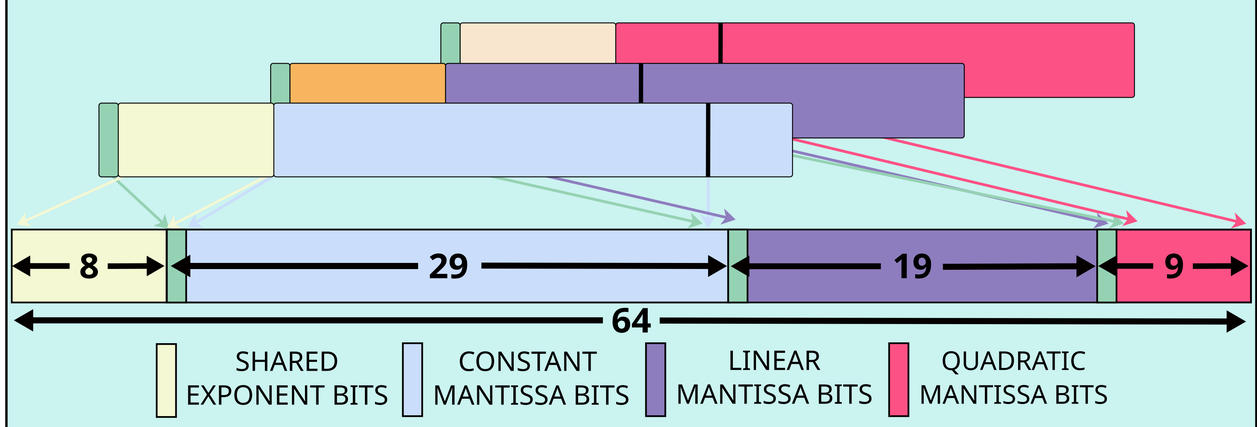






Error convergence of different data representations in DG methods.
SBFP: 29,19 and 8 bits for constant, linear and quadratic terms.
ASBFP (29, 19, 8): 29, 19, and 8 bits for constant, linear, and quadratic terms.
ASBFP (25, 20, 11): 25, 20, and 11 bits for constant, linear, and quadratic terms.

Spectral Block Floating Point (SBFP)



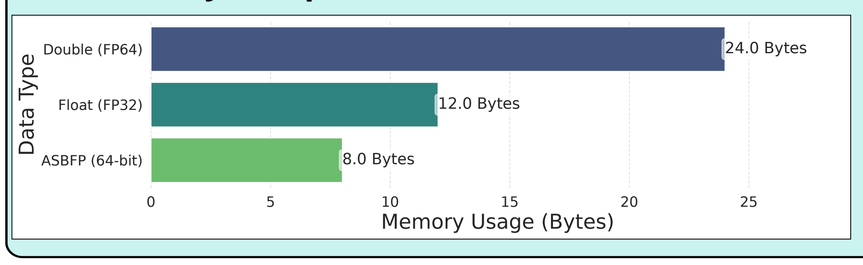
Core Design Principles:

Shared Exponent	Compact Encoding 64 bits	Cache-Friendly Single Word

Arithmetics

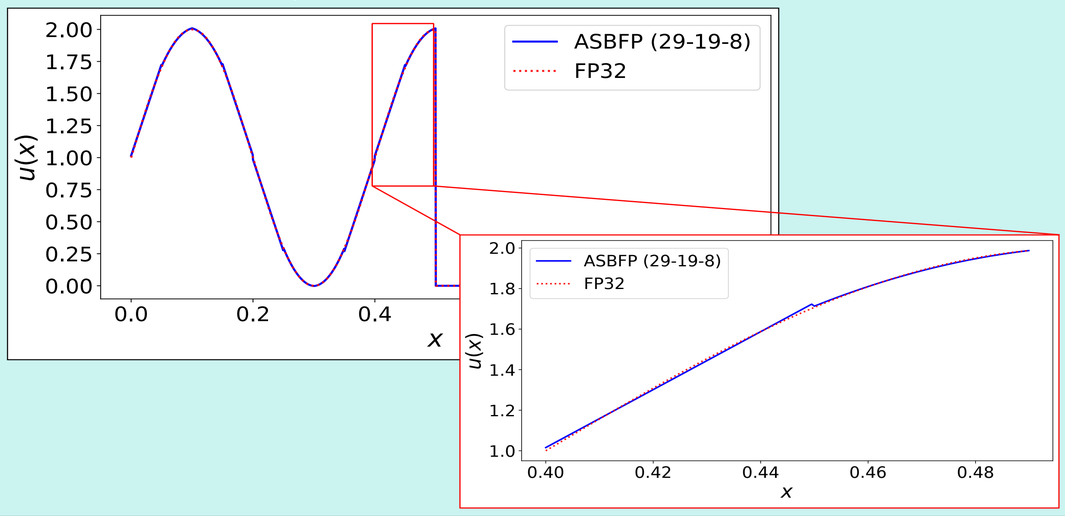
Decode → IEEE 754 Representation → Operate → Re-encode

Memory Comparison: FP64 vs FP32 vs BSFP



Discontinuous Initial Condition

$$u(x,t) = \begin{cases} \sin(5\pi x - t) + 1, & \text{if } x < 0.5 \\ 0, & \text{if } x \geq 0.5 \end{cases} \quad \text{for } x \in [0,1]$$



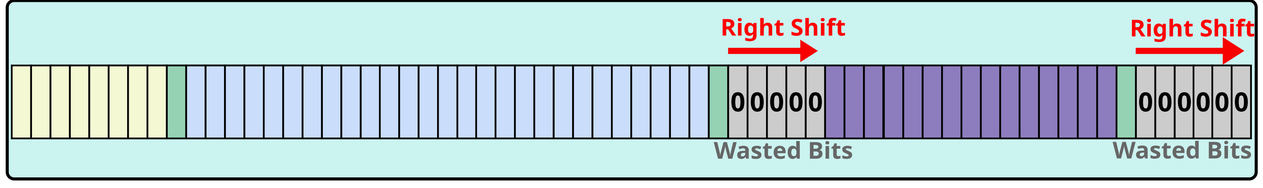
Key Takeaways

• SBFP:
- Reduces overhead but causes bit loss due to shared exponent allignment.

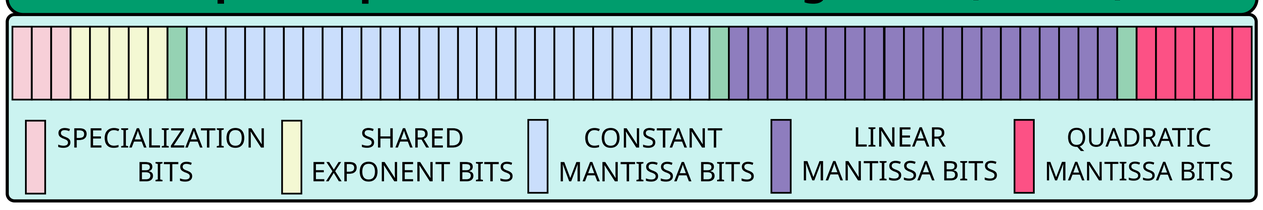
• ASBFP:
- adapts encoding via specialisation → better precision & memory use.

• Proven Performance:
- High accuracy in 1D DG — beats FP32 at fine mesh.

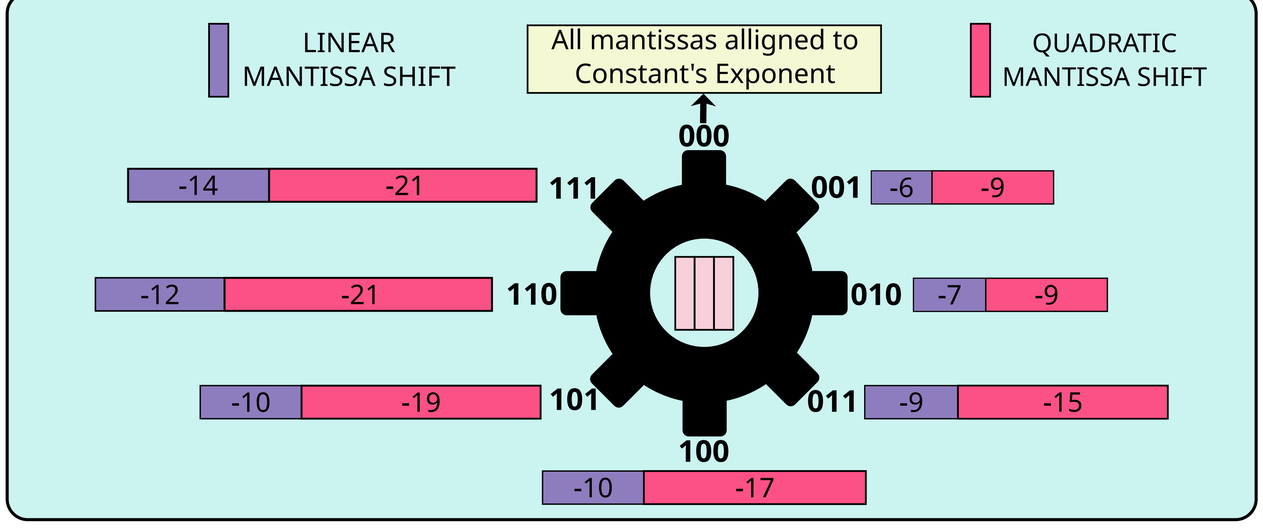
SBFP Drawback: Excessive Bit Shifts from Shared Exponent Alignment



Adaptive Spectral Block Floating Point (ASBFP)



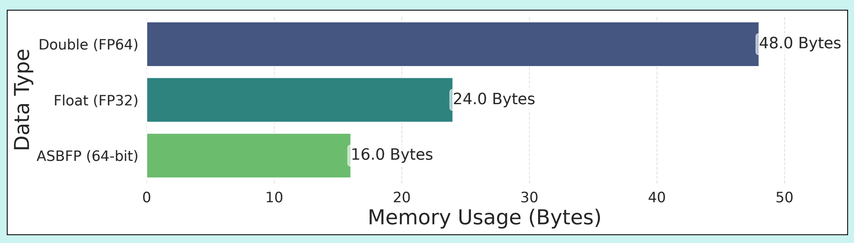
ASBFP Specialisation Gear: Adaptive Exponent Strategies



Future Work

• **Discontinuities**
- Extend methods to handle non-smooth solutions (e.g., shocks)

• **2D Extension**
- Apply ASBFP to 2D Discontinuous Galerkin (DG) methods



• **Broader PDEs**
- Test on:
  • Shallow Water
  • Euler Equations

References

1-L. Einkemmer, "A mixed precision semi-Lagrangian algorithm and its performance on accelerators," 2016 International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, Austria, 2016, pp. 74-80, doi: 10.1109/HPCSim.2016.7568318.